


---

**Monograph:**
Brazier, J., Rowen, D., Yang, Y. et al. (1 more author) (2009) Using rank and discrete choice data to estimate health state utility values on the QALY scale. Discussion Paper. (Unpublished)

HEDS Discussion Paper 09/10

---


# HEDS Discussion Paper 09/10


ScHARR
School of Health and
Related Research

The University Of Sheffield.

ScHARR

# Health Economics and Decision Science Discussion Paper Series

No. 09/10

# Using rank and discrete choice data to estimate health state utility values on the QALY scale

John Brazier[a] , Donna Rowen[a]*, Yaling Yang[a], and Aki Tsuchiya [a,b]

[a] Health Economics and Decision Science, University of Sheffield, Regent Court, 30Regent Street, Sheffield, S1 4DA, UK
[b] Department of Economics, University of Sheffield, 9 Mappin Street, Sheffield, S1 4DT, UK
* Correspondence to: Donna Rowen, Health Economics and Decision Science,University of Sheffield, Regent Court, 30 Regent Street, Sheffield, S1 4DA, UK.
Telephone: +44114 222 0728.
Fax: +44114 272 4095.
Email:d.rowen@sheffield.ac.uk

# USING RANK AND DISCRETE CHOICE DATA TO ESTIMATE HEALTH STATE UTILITY VALUES ON THE QALY SCALE

John Brazier PhD[1], Donna Rowen PhD[1], Yaling Yang MSc[1] and Aki Tsuchiya PhD[1,2]

1: Health Economics and Decision Science, School of Health and Related Research, University of Sheffield
2: Department of Economics, University of Sheffield

Corresponding author: Donna Rowen, Health Economics and Decision Science, School of Health and Related Research (ScHARR), University of Sheffield, Regent Court, 30 Regent Street, Sheffield S1 4DA. d.rowen@sheffield.ac.uk
Tel: +44(0)114 222 0728. Fax: +44(0)114 272 4095.

Financial support: The studies reported in this paper were funded by Novartis (AQL-5D) and Pfizer Inc. (OAB-5D).
Keywords: Ranking, discrete choice experiment, preference-based measures, QALYs
Running title: Estimating utilities using rank and DCE data

**Abstract**

**Objective:** Recent years has seen increasing interest in the use of ordinal methods to elicit health state utility values as an alternative to conventional methods such as standard gamble and time trade-off. However in order to use these health state values in cost effectiveness analysis using cost per quality adjusted life year (QALY) analysis these values must be anchored on the full health-dead scale. This study addresses this challenge and examines how rank and discrete choice experiment data can be used to elicit health state utility values anchored on the full health-dead scale and compares the results to time trade-off (TTO) results.

**Methods:** Two valuation studies were conducted using identical methods for two health state classification systems; asthma and overactive bladder. Each valuation study involved interviews of 300 members of the general population using ranking and TTO plus a postal survey using discrete choice experiment sent to all consenting interviewees and a 'cold' sample of the general population who were not interviewed.

**Results:** Overall DCE produced different results to ranking and time trade-off whereas ranking produced similar results to TTO in one study, but not the other.

**Conclusions:** Ordinal methods offer a promising alternative to conventional cardinal methods of standard gamble and TTO. However the results do not appear to be robust across different health state classification systems and potentially different medical conditions. There remains a large and important research agenda to address.

A**cknowledgements**

We would like to thank the Centre for Research and Evaluation at Sheffield Hallam University for conducting the interviews. We are grateful to members of HESG for their comments on the paper. The studies reported in this paper were funded by Novartis (ALQ-5D) and Pfizer Inc. (OAB-5D). The usual disclaimer applies.

**Introduction**

The status of preference-based measures of health for generating Quality Adjusted Life Years (QALYs) was considerably enhanced by the recommendations of the U.S. Public Health Service Panel on Cost-Effectiveness in Health and Medicine to use them in economic evaluation [1]. The use of preference-based measures has grown considerably over the last decade with the increasing use of economic evaluation to inform health policy, for example through the establishment of bodies such as the National Institute of Health and Clinical Excellence in England and Wales [2].

To be a preference-based measure it has been suggested that the health state valuation technique must be choice-based [1,2,3]. The two choice-based techniques most commonly used to value preference-based measures are the cardinal methods of standard gamble (SG) and time trade-off (TTO) [4,5,6]. There are concerns about these cardinal methods because they are likely to be affected by factors other than a respondent's preference for the state, such as risk aversion in the case of standard gamble or time preference and aversion to losses for TTO [7]. Furthermore, these tasks are cognitively complex and respondents might have some difficulty with them, particularly those in vulnerable groups such as the very elderly or children. For these reasons there has been increasing interest in using ordinal tasks that require the respondent to rank one or more states [8,9,10] and in discrete choice experiments (DCE) involving pairwise comparisons [11,12,13].

The ability to derive cardinal health state values from ordinal information comes from the assumption that a respondent's selection over a set of states will be related to a latent variable. It allows for the fact that individuals make errors of judgement and sometimes may choose the health state with a lower value. The proportion of occasions on which such an error is made is related to the distance between values of the states in terms of the latent variable. There will be more agreement in preferences the further apart the values for two states. This has been the basis for the more general use of discrete choice experiments. By making additional assumptions it is possible to 'explode' ranking data into discrete choice data, whereby the ordering of X states is essentially seen as a sequence of discrete choices.

A key problem in using ordinal methods has been how to anchor the values estimated by logistic models onto the full health-dead scale required for generating QALYs, anchoring full health at one and dead at zero. If the preference weights do

not produce utility values on the full health-dead scale they cannot be used in economic evaluation using cost per QALY analysis. This paper addresses the problem of anchoring onto the full health-dead scale in the context of two valuation studies, one for an asthma-specific measure and the other for an overactive bladder-specific measure. The paper begins by presenting an overview of the theory underlying the ordinal methods. The methods and results of the valuation studies are presented, including a comparison of results using ranking, DCE and TTO on the same full health-dead scale. Results from the DCE data obtained from a sample that had previously been interviewed are also compared to those obtained from a 'cold' sample that had not previously been interviewed. The implications of this study for further work are considered in the discussion.

## Theoretical basis for deriving cardinal values for health states from ordinal information

The idea of obtaining cardinal values from ordinal data first came from the work of Thurstone [14] who proposed the 'law of comparative judgement'. This was recognised [15] as offering a method for deriving cardinal preferences for health states from rank preference data and later implemented using the sleep dimension of the Nottingham Health Profile [8] and more recently the EQ-5D classification [16].

Thurstone's approach has been modified in a number of ways, including the application of a logistic function [17,18] as a means of modelling the latent utility function from ordinal data. Another important modification in this context is that in modelling a population level latent utility function from individual rank data, the error is being characterised in terms of the deviation of the individuals' preferences from the population preferences; i.e. variation in individual preferences within a population is considered analogous to Thurstone's individual level perceptual error. To use rank data the assumption of independence from irrelevant alternatives (IIA) is required in order to explode the rank data into a series of pairwise choices. This assumes that the ordering of a pair of states does not depend on the other states being considered.

Recently conditional logistic regression models were applied to the rank data collected as part of the UK valuation of the EQ-5D [9], SF-6D and HUI2 [10]. The rank model of health states alone does not produce utilities on the full health-dead scale necessary for use in generating QALYs, as it does not enable the anchoring of the values to 0 for dead. For this reason, the values generated by the logit model are

transformed onto the full health-dead scale needed to generate QALYs. One method involves normalising the coefficients using the mean TTO value for the worst state defined by the classification system [9]. An alternative approach is to include the state 'dead' in the ranking exercise and normalise the regression coefficients so that 'dead' achieves a predicted value of zero [10].

DCE is a widely used tool in health economics for eliciting values, but has so far had limited use for eliciting values for preference-based measures of health used to derive QALYs. A limited number of studies have used DCE to value health states for their own sake [11,19,20,21,12,13], but none have anchored their results onto the full health-dead scale required for generating QALYs. One study attempts a partial solution by normalising the DCE results using the estimated TTO value for the worst possible state [12]. The studies presented in this paper are the first attempt to undertake a normalisation of DCE results around dead without the use of cardinal values obtained from external sources. Here we include the state 'dead' in the DCE and use this directly estimated parameter to rescale the regression coefficients. We compare the results to those obtained using the alternative approach of normalising using the estimated TTO value for worst state [12].

## Methods

### The health state classifications

#### Asthma specific-measure

The AQL-5D is a 5-dimension health state classification system [22] developed from the Asthma Quality of Life Questionnaire, AQLQ [23]. The dimensions of AQL-5D are: concern about asthma, shortness of breath, weather and pollution stimuli, sleep impact and activity limitations (Table 1). The health state classification system has 5 dimensions each with 5 levels of severity, with level 1 denoting no problems and level 5 indicating extreme problems. By selecting one level for each dimension it is possible to define 3125 health states.

#### Overactive bladder-specific measure

The OAB-5D is a 5-dimension health state classification [24] developed from the overactive bladder instrument, OABq [25]. The dimensions of the OAB-5D are: urge, urine loss, sleep, coping and concern (Table 2). The health state classification system has the same structure as the AQL-5D, defining a total of 3125 health states.

### Interview

Two valuation surveys were conducted, one for each health state classification. These surveys were identical in design in everyway, apart from using different health state classifications to define the health state descriptions. Sample sizes differed for the DCE due to funding constraints. The surveys elicited values for a selection of states (AQL-5D/OAB-5D) from a representative sample of 300 members of the general public each. Adults who consented to participate were interviewed in their own home by an experienced interviewer trained by the authors of this paper. Respondents were asked to complete the health state classification questionnaire for themselves to help familiarise them with it. The first valuation task was to rank 7 intermediate states, full health (health state 11111), worst state defined by the health state classification ('pits' state 55555), and immediate death. The ranking task has been used in the past in valuation studies for the EQ-5D [4] and SF-6D [5] and has conventionally been seen as a warm up task to the main cardinal task.

The next task was to value the 7 intermediate states and 'pits', with an upper anchor of full health using TTO. The survey used the TTO-prop method developed by the York Measurement and Valuation Health Group, which uses a 'time board' as a visual aid [26]. Respondents were then asked a series of socio-demographic questions. Finally, they were asked about their willingness to participate in a postal survey (described below).

The selection of health states for the interviews was determined by the specification of the model to be estimated. In this study, 98 health states, and the worst state (to be repeated across the design) were selected out of the 3125 possible health states described by the classification system. The selection was on the basis of a balanced design, which ensured that any dimension-level (level $\lambda$ of dimension $\delta$) had an equal chance of being combined with all levels of the other dimensions. These 98 states were stratified into severity groups based on their total level score across the dimensions (simply the sum of the levels), and then randomly allocated into 14 blocks, so that each block has 7 health states. This procedure ensured that each respondent, who was allocated one of the 14 blocks, received a set of states balanced in terms of severity and that each state is valued the same number of times except the worst possible state, the 'pits' state, which is valued by all respondents.

**Postal surveys**

A DCE questionnaire was mailed to interviewees who had consented to the postal survey approximately four weeks after the interviews (the 'warm' sample). The same

questionnaire was mailed out to a separate sample of the general public who had not been interviewed (the 'cold' sample'). The cold sample size was determined by funding constraints. Respondents were asked to complete the health state classification questionnaire for themselves to help familiarise them with it. Respondents were asked to indicate which state they preferred for an example pair of states and then for 8 pairs of states (see example question in Table 3). Finally respondents were asked a series of socio-demographic questions. Reminders were sent to all non-responders approximately four weeks after the initial questionnaire was sent.

The large number of states defined by the classification systems of each measure mean it is infeasible to value all states. States were selected for the postal DCE using an application of a specially developed programme in the statistical package SAS [27]. The programme obtains an optimal statistical design for DCE based on level balance, orthogonality, minimal overlap and utility balance. This reduces the number of pairwise comparisons to a manageable number. The programme produced 12 pairwise comparisons from the AQL-5D and OAB-5D, and these were randomly allocated to two versions of the questionnaire with 6 pairwise choices each. Two additional pairwise comparisons were included of two poor health states each compared to 'immediate death', and these were common across all versions of the questionnaire. No other states or pairwise comparisons were included in each version of the questionnaire. Only one pairwise comparison involves a logically consistent choice where one state has better health for every dimension.

**Modelling health state values**

**Time trade-off**

The data from the TTO valuation exercise was analysed using a one way error components random effects model which takes account of variation both within and between respondents [5]. The standard model is defined as:

$$y_{ij} = f(\boldsymbol{\beta}\mathbf{x}_{ij}) + \varepsilon_{ij} \quad (1)$$

Where $i$=1,2 …$n$ represent individual health state values and $j$=1,2…$m$ represents respondents. The dependent variable $y_{ij}$ is the disvalue (1–TTO value) for health state $i$ valued by respondent $j$ and $\mathbf{x}_{\partial\lambda}$ is a vector of dummy explanatory variables for

each level $\lambda$ of dimension $\partial$ of the health state classification. Level $\lambda = 1$ acts as a baseline for each dimension. $\varepsilon_{ij}$ is the error term which is subdivided as follows:

$$\varepsilon_{ij} = u_j + e_{ij} \quad (2)$$

Where $u_j$ is respondent specific variation and $e_{ij}$ is an error term for the $i$th health state valuation of the $j$th individual, and this is assumed to be random across observations. Details of other models run on the TTO data are available elsewhere for both AQL-5D [28] and OAB-5D [29].

**Ranking**

The rank ordered logit model was used to analyse the ranking data (a modelling approach also referred to as the conditional logit model [30]). It states that respondent $j$ has a latent utility function for state $i$, $U_{ij}$ and given the choice of two states $i$ and $k$, the respondent will choose state $j$ over state $k$ if $U_{ij} > U_{ik}$.

The expected value of each unobserved utility was assumed to be a linear function of the categorical levels on the dimensions of the health state classification. Following the approach taken elsewhere [9,10] the general model specification for each individual $j$'s cardinal utility function for state $i$ is $U_{ij} = \mu_j + \varepsilon_{ij}$ where $\mu_j$ is representative of the tastes of the population and $\varepsilon_{ij}$ represents the particular tastes of the individual. If the error term $\epsilon$ has an extreme value distribution, then the odds of choosing state over state $k$ are $exp\{\mu_j - \mu_k\}$.

The general model specification for analysis of the ranking data is:

$$U_{ij} = \boldsymbol{\beta}\mathbf{x}_{ij} + \Phi D_j + u_{ij} \quad (3)$$

where $U$ represents utility; $j$=1,2,…$n$ represents respondents and $i$ = 1,2,….$m$ represents health states. The functional form is assumed to be linear. The vector of dummies is as defined for equation (1), with the addition of a dummy variable for the state dead. For all health states other than dead D = 0. In order to anchor onto the full health-dead scale the coefficients relating to the levels of each dimension are normalised by dividing each level coefficient by the coefficient relating to dead [9,10].

**Discrete Choice Experiment**

The data from the DCE surveys were analysed using a random effects probit model, which takes account of the repeated measurement aspect of the data (whereby multiple responses are obtained from the same individual). Again an additive specification is used as specified by equation (3). The coefficients were normalised in the same way as the rank data by dividing each level coefficient by the coefficient relating to dead. Models are also estimated for the 'warm' sample that was previously interviewed and the 'cold' sample that were not. Finally the DCE data is also modelled using an existing approach in the literature [12]. This approach estimates a random effects probit model using the DCE data excluding the pairwise comparisons involving 'dead'. The coefficients are normalised onto the full health-dead scale using the estimated TTO value of the worst state.

**Comparison of models**

The three models are compared. There is no reason why rank or DCE models should produce the same results as the TTO model, although it could be thought that Rank and DCE may produce similar results as the use of the rank-ordered logit model means that the rank data is viewed as a series of pairwise comparisons.

Models can be compared in terms of the sign and ordering of their coefficients. The sign of the coefficients on the levels of each dimension are expected to be negative since they are all worse than the baseline (i.e. level 1). Furthermore, the levels in each dimension have a logical ordering, whereby more severe levels should have larger decrements. The number of inconsistencies between significant coefficients is compared between the models. For interest, we examine the relationship between model predictions and TTO observed values including the mean absolute difference, the root mean square of the difference and the proportions of differences greater than 0.05 and 0.1. Finally the pattern of the predictions is compared.

## Results

### The interview respondents

Three hundred and seven members of the public (response rate of 40%) in South Yorkshire (UK) were interviewed in the AQL-5D survey and 311 people interviewed in the OAB-5D survey (response rate of 26.7%). Table 4 shows that the two samples were very similar in terms of their socio-demographic composition. Among the respondents to the AQL-5D survey, 53 (17.3%) had asthma and in the OAB-5D survey 27 (8.7%) reported experiencing symptoms of urge and 18 (5.8%) reported

urine loss for at least some of the time. Overall self-reported health status using EQ-5D [4] was very close to the UK EQ-5D norms of 0.85 for females and 0.86 for males [31]. Two hundred and sixty three people responded to the AQL-5D postal survey and 402 people responded to the OAB-5D postal survey. Table 4 shows that the socio-demographic composition of the postal samples are similar to the interview samples, but the OAB-5D postal survey has a larger proportion of respondents over 65 years of age and a higher proportion of females. Overall the AQL-5D samples have lower mean EQ-5D scores.

**The data set**

**AQL-5D**

There were 2455 TTO health state valuations generated by the 307 respondents from the interviews and 3041 states ranked by the respondents at their interview. The average number of TTO valuations per intermediate health state was 22 (range from 19 to 22) and the 'pits' state (AQL-5D state 55555) was valued by every respondent (n=307). Mean TTO health state values ranged from 0.39 to 0.94 and generally have fairly large standard deviations (around 0.2 to 0.4). The distribution of the values was negatively skewed.

There were 168 DCE questionnaires returned out of the 308 who had been interviewed (55%) generating 1336 observed pairwise comparisons. In total 95 DCE questionnaires were returned in the cold survey (a 23% return rate) generating 741 pairwise comparisons.

**OABq**

There were 2487 health state values generated by the 311 respondents and 3040 states ranked. Each intermediate health state was valued 22 times using TTO (range from 17 to 29) and the worst possible state (OAB-5D 55555) was valued 310 times using TTO (one missing value). Mean TTO health state values ranged from 0.56 for the worst possible state, to 0.91 for state 13321, with an average standard deviation of 0.28.

The warm survey had 133 returned DCE questionnaires (response rate 44%) generating 1050 pairwise comparisons. The cold survey resulted in 268 being returned (response rate 27%) generating 2059 comparisons.

**Modelling**

**AQL-5D**

The TTO model and transformed rank and DCE models are presented in Table 5. The TTO model produced the expected negative coefficients for all statistically significant coefficients and the ordering of coefficients was consistent with the dimension levels of the AQL-5D. Three coefficients were positive but statistically insignificant. The rank model produced all negative coefficients and no inconsistencies for all significant coefficients. In comparison to the TTO and rank models the DCE models normalised using the dead coefficient have a higher number of positive coefficients and inconsistencies. The warm DCE model produced five positive coefficients, none of which were statistically significant, and one inconsistency amongst statistically significant coefficients. The cold model had one positive coefficient that was not statistically significant and no inconsistencies between significant coefficients. The DCE models for the pooled data (i.e. warm plus cold) produced three positive coefficients, one of which is statistically significant, and one inconsistency between significant coefficients. The weather dimension seemed to cause most difficulty for the DCE models, with a suggestion that the levels of this dimension do not conform to the suggested ordering. The DCE model using the estimated TTO value for the worst state has four positive coefficients, one of which is statistically significant, and one inconsistency between significant coefficients.

The size of the dimension level coefficients of the rank and TTO models are quite similar and follow an orderly pattern against the levels of the AQL-5D. The DCE model for the pooled data set reveals some marked differences. The most noticeable differences lie at the lower end of concern, short of breath, pollution and the upper ends of sleep and activity. Level 2 for the dimensions of concern, breath and pollution are all positive and in the wrong direction, quite markedly so for pollution. Sleep and activity have coefficients with the right sign, but they are much larger for levels 4 and 5.

The similarity of the rank and TTO models can be seen in the plot of predicted health state values against observed mean TTO values in Figure 1. Mean absolute differences from observed TTO are 0.056 and 0.061 for the TTO and rank models respectively, with mean differences of around zero. By contrast, the DCE predictions follow different paths depending on the normalisation method used. The DCE model that rescaled coefficients using the rank method tended to have health state predicted values that were higher than observed TTO whereas the DCE model that rescaled coefficients using estimated TTO value for worst state tended to have health

state values lower than observed TTO values. The results from the DCE model that rescaled coefficients using the estimated TTO value for worst state is more similar to the TTO model estimates, as expected due to the method of normalisation. Differences are observed between the mean values for the worst AQL-5D health state of 0.390 for observed TTO and predictions of 0.431 for TTO, 0.434 for rank data and 0.154 for predictions from pooled DCE data normalised using the dead coefficient.

**OAB-5D**

The OAB-5D results are presented in Table 6. Overall the models were broadly consistent with the ordinality of the OAB-5D. All the coefficients in the TTO model were negative and most significant. There were inconsistencies between significant coefficients in 3 cases, but their magnitudes were 0.02 or less. The ranking data produced negative coefficients and all but one were statistically significant and no inconsistencies between significant coefficients. The DCE model using the warm sample had five positive coefficients, but none were significant. All DCE models normalised using the dead coefficient have five positive coefficients, one of which is statistically significant (coping level 2) and two inconsistencies amongst the significant coefficients.

The TTO model does not predict observed TTO as well as for the AQL-5D as indicated by mean absolute deviation (MAD) and mean error in Tables 5 and 6. Ranking predictions also do not agree with TTO as closely as for the AQL-5D survey and tended to have predicted health state values that are higher than observed TTO values. As for the AQL-5D survey, the DCE predictions have a larger scale range (0.249 to 1.00 compared to 0.623 to 1.0 for TTO and 0.436 to 1.0 for ranking). Again the DCE models have different results depending on the method of normalisation. Again the model using the dead coefficient to rescale coefficients tended to have predicted health state values higher than observed TTO, whereas the model using the estimated TTO value of worst state to rescale tended to have predicted health state values lower than observed TTO.

**Discussion**

This study has shown how DCE and rank data can be used to generate health state values on the full health-dead scale required to generate QALYs. As would be expected, the TTO model best predicted TTO observed values, but then there is no reason to expect rank and DCE data to produce the same values. Perhaps more

surprising is the way the rank model coefficients were actually very similar to the TTO in the AQL-5D survey, but less so in the OAB-5D survey. In both surveys the DCE model was the most different from the other methods, and the model normalising coefficients using the dead coefficient produced a larger range of values.

In modelling, rank data are essentially treated as data series of pairwise comparisons, and aside from the IIA assumptions, are otherwise the same. It is therefore interesting to find that they do not produce the same values. This may suggest that the rank and DCE tasks generate different data, which may have implications for the IIA assumption used in rank data. However it may also reflect the fact that the ranking task preceded the TTO in the same interview, whereas the DCE data were collected via a postal survey. Furthermore different states are valued in the rank and DCE tasks.

For the DCE surveys, despite the fact that one sample had been interviewed previously and the other had not, there seem to be little obvious difference in terms of the coefficients. Although the sample sizes are small for the 'cold' and 'warm' samples, particularly for the cold AQL-5D sample. This suggests that it may be possible to obtain DCE data to value health states without prior interview. This would be considerably cheaper, but postal surveys are usually associated with lower response rates and this was true for the AQL-5D survey. For researchers seeking to use DCE without other methods, it may still be preferable to approach respondents directly in their own home to ensure a more representative sample.

The pooled DCE models using different methods to rescale onto the full health-dead scale produce noticeably different coefficients and different ranges or predicted values. As expected the model normalising coefficients using the estimated TTO value of worst state is more similar to the observed TTO values and the TTO model. Overall the results suggest that DCE and TTO produce different results, and the use of TTO data to rescale DCE coefficients rather than using data collected using a DCE alone produces different results. This should be recognised in the future design of DCE surveys to obtain health state values.

The DCE were really 'add-ons' to a study that was mainly designed to provide TTO valuations of the AQL-5D and OAB-5D. Using a postal method for DCE, for example, may have compromised the quality of the data and it certainly resulted in a lower response rate. Perhaps more importantly, the recommended approach for state

selection and design for DCE experiments continually evolves [32], and our study may have benefited from recent improvements in DCE design.

The DCE models based on the warm and cold samples seem to have similar coefficients and so were pooled to focus on the main comparisons with TTO and rank and the existing alternative approach used to anchor values onto the full-health to dead scale [12]. However, the pooled data should be treated with some caution. Further analysis did find some difference between the samples. A dummy variable for 'cold' was significant in both surveys with values of -0.06 for AQL-5D and -0.045 for the OAB-5D on the full health-dead scale. These results suggest the cold sample gave slightly lower values than the sample that had previously been interviewed, though this difference is not sufficiently large to alter the main findings comparing the different valuation methods.

There are concerns with the types of models estimated here since they make restrictive distributional assumptions about the coefficients. Of particular concern is that some orderings are logically determined. For example, suppose there is a health state pair: *j* and *k*, and $\mu_j - \mu_k = X$, say 0.2, on the latent variable scale standardised to 1 for full health and 0 for dead. The current approach to modelling ordinal data assumes that any two states that are apart from each other by *X* will have the same proportion of respondent's incorrectly ranking *j* over *k*. However, it is reasonable to assume that the probability of error will not only be a function of how apart the two states are, but also whether or not the two states have a logically determined ordering. Suppose there are two sets of health state pairs that are apart by *X*, where pair 1 has no logically determined ordering (e.g. 11122 and 33111) whereas pair 2 has a logically determined ordering (e.g. 11122 and 11133). It is reasonable to expect that the proportion of responses that rank *j* over *k* will be different across pair 1 and pair 2. This becomes particularly problematic when one of the states is full health or the worst state. This means that the structure of the error term in equation (3) needs to be more sophisticated than it currently is. There are now more advanced econometric modelling techniques known as mixed logit models [33] that should be explored with both these data sets. This would also overcome the IIA assumption underlying the way rank data are being analysed.

A key methodological innovation presented in this paper has been to include dead as a state in the pairwise choices and then to use this to anchor the values generated by the logistic models. Another way to achieve this anchoring would be to include

survival as a separate attribute. However, this would require a far larger and more complex design, since survival has a multiplicative relationship to health related quality of life in the QALY model. The disadvantage with including dead as a state arises from the fact that many respondents may not regard any state defined by the classification as worse than being dead and so effectively not be willing to trade. This is likely to be more of a problem for milder descriptive systems. For these studies, a sufficient proportion of respondents were willing to make a trade, so that at the aggregate level it has been possible to estimate a societal value for the state of being dead compared to the health states defined by the health state classification.

**Conclusion**

This study has shown how rank and DCE data can be used to generate health state values using the QALY scale. It proposes a new method for doing this that includes dead in the DCE exercises in order to anchor the health state values. While ordinal methods may offer a promising alternative to conventional cardinal methods of SG and TTO, there is a large and important research agenda to address.

**References**

[1] Gold MR, Siegel JE, Russell LB, Weinstein MC. Cost-Effectiveness in Health and Medicine. Oxford: Oxford University Press, 1996.

[2] NICE (National Institute for Health and Clinical Excellence). Guide to the methods of technology appraisal. NICE: London, 2008.

[3] Drummond MF, Sculpher M, O'Brien B, et al. Methods for the economic evaluation of health care programmes. Oxford: Oxford Medical Publications, 2005.

[4] Dolan P. Modelling valuation for Euroqol health states. Med Care 1997;35:351-363.

[5] Brazier J, Roberts J, Deverill M. The estimation of a preference based single index measure for health from the SF-36. J Health Econ 2002;21:271-292.

[6] Feeny D, Furlong W, Torrance G, et al. Multiattribute and single attribute utility functions for the Health Utilities Index Mark 3 system. Med Care 2002;40:113-128.

[7] Bleichrodt H. A new explanation for the difference between time trade-off utilities and standard gamble utilities. Health Econ 2002;11:447-456.

[8] Kind P. A comparison of two models for scaling health indicators. Int J Epidemiol 1982;11:271–275.

[9] Salomon JA (2003). Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. *Popul Health Metr.* 2003;1:12.

[10] McCabe C, Brazier J, Gilks P, et al. Using rank data to estimate health state utility models. J Health Econ 2006;25:418-431.

[11] Burr JM, Kilonzo M, Vale L, et al. Developing a Preference-Based Glaucoma Utility Index Using a Discrete Choice Experiment. Optom Vis Sci 2007;84:797-808.

[12] Ratcliffe J, Brazier J, Tsuchiya A, et al. Using DCE and ranking data to estimate cardinal values for health states for deriving a preference-based single index from the sexual quality of life questionnaire. Health Econ 2009; Forthcoming.

[13] Ryan M, Netten A, Skatun D, et al. Using discrete choice experiments to estimate a preference-based measure of outcome – an application to social care for older people. J Health Econ 2006;25:927-944.

[14] Thurstone, LL. A law of comparative judgement. Psychol Rev 1927;34:273-286

[15] Fanshel S, Bush JW. A health status index and its application to health services outcomes. Operations Research 1970;18:1021-1066.

[16] Kind P. Applying paired comparisons models to EQ-5D valuations – deriving TTO utilities from ordinal preferences data. In: Kind P, Brooks R, Rabin R eds., EQ-5D concepts and methods: a developmental history. Netherlands, Springer, 2005.

[17] Luce RD. Individual choice behavior: a theoretical analysis. New York: John Wiley & Sons, Inc., 1959.

[18] McFadden D. Conditional logit analysis of qualitative choice behavior. In Zarembka P ed., Frontiers in econometrics. New York; Academic Press, 1974.

[19] Hakim Z, Pathak DS. Modelling the EuroQol data: a comparison of discrete choice conjoint and conditional preference modelling. Health Econ 1999;8:103-116.

[20] Johnson R, Banzhaf M Desvousges W. Willingness to pay for improved respiratory and cardiovascular health: a multiple-format, stated preference approach. Health Econ 2000;9:295–317.

[21] Osman LM, McKenzie L, Cairns J, et al. Patient weighting of importance of asthma symptoms. Thorax 2001;56:138-42.

[22] Young T, Yang Y, Brazier J, et al. The use of Rasch analysis as a tool in the construction of a preference based measure: the case of AQLQ. Health Economics and Decision Science Discussion Paper 07/01. ScHARR, University of Sheffield, 2007. http://www.sheffield.ac.uk/scharr/sections/heds/discussion.html

[23] Juniper EF, Guyatt GH, Ferrie PJ, et al. Measuring quality of life in asthma. American Review of Respiratory Disease 1993;147:832-838.

[24] Young T, Yang Y, Brazier J, et al. The first stage of developing preference-based measures: constructing a health-state classification using Rasch analysis. Qual Life Res 2009;18:253-265.

[25] Coyne K, Revicki D, Hunt T, et al. Psychometric validation of an overactive bladder symptom and health related quality of life questionnaire: The OAB-q. Qual Life Res 2002;11:563-574.

[26] MVH Group. The measurement and valuation of health: Final report on the modelling of valuation tariffs. Centre for Health Economics, University of York, 1995.

[27] Huber J, Zwerina K. The importance of utility balance in efficient choice designs. Journal of Marketing Research 1996;33:307-317.

[28] Yang Y, Tsuchiya A, Brazier J, et al. Estimating a preference-based single index from the Asthma Quality of Life Questionnaire (AQLQ). Health Economics and Decision Science Discussion Paper 07/02. ScHARR, University of Sheffield, 2007. http://www.sheffield.ac.uk/scharr/sections/heds/discussion.html

[29] Yang Y, Brazier JE, Tsuchyia A, et al. Estimating a preference-based index from the Over Active Bladder questionnaire. Value Health 2009;12:159-166.

[30] Brazier J, Ratcliffe J, Salomon J, et al. The measurement and valuation of health benefits for economic evaluation. Oxford University Press, Oxford, 2007.

[31] Kind P, Harman G, Macran S. UK population norms for EQ-5D. Centre for Health Economics Discussion Series, University of York, 1999.

[32] Louviere JJ. What you don't know might hurt you: some unresolved issues in the design and analysis of discrete choice experiments. Environmental and Resource Economics 2006;34:173–88.

[33] Train K. Discrete choice methods with simulation. Cambridge University Press. Cambridge, 2003.

**Table 1** Asthma quality of life classification (AQL-5D)

**Concern**

1. Feel concerned about having asthma none of the time
2. Feel concerned about having asthma a little or hardly any of the time
3. Feel concerned about having asthma some of the time
4. Feel concerned about having asthma most of the time
5. Feel concerned about having asthma all of the time

**Short of breath**

1. Feel short of breath as a result of asthma none of the time
2. Feel short of breath as a result of asthma a little or hardly any of the time
3. Feel short of breath as a result of asthma some of the time
4. Feel short of breath as a result of asthma most of the time
5. Feel short of breath as a result of asthma all of the time

**Weather and pollution**

1. Experience asthma symptoms as a result of air pollution none of the time
2. Experience asthma symptoms as a result of air pollution a little or hardly any of the time
3. Experience asthma symptoms as a result of air pollution some of the time
4. Experience asthma symptoms as a result of air pollution most of the time
5. Experience asthma symptoms as a result of air pollution all of the time

**Sleep**

1. Asthma interferes with getting a good night's sleep none of the time
2. Asthma interferes with getting a good night's sleep a little or hardly any of the time
3. Asthma interferes with getting a good night's sleep some of the time
4. Asthma interferes with getting a good night's sleep most of the time
5. Asthma interferes with getting a good night's sleep all of the time

**Activities**

1. Overall, not at all limited with all the activities done
2. Overall, a little limitation with all the activities done
3. Overall, moderate or some limitation with all the activities done
4. Overall, extremely or very limited with all the activities done
5. Overall, totally limited with all the activities done

**Table 2** Overactive bladder quality of life classification system (OAB-5D)

**Urge**
1. Not at all bothered by an uncomfortable urge to urinate
2. Bothered by an uncomfortable urge to urinate a little bit or somewhat
3. Bothered by an uncomfortable urge to urinate quite a bit
4. Bothered by an uncomfortable urge to urinate a great deal
5. Bothered by an uncomfortable urge to urinate a very great deal

**Urine loss**
1. Not at all bothered by urine loss associated with a strong desire to urinate
2. Bothered by urine loss associated with a strong desire to urinate a little bit or somewhat
3. Bothered by urine loss associated with a strong desire to urinate quite a bit
4. Bothered by urine loss associated with a strong desire to urinate a great deal
5. Bothered by urine loss associated with a strong desire to urinate a very great deal

**Sleep**
1. Bladder symptoms interfered with your ability to get a good night's rest none of the time
2. Bladder symptoms interfered with your ability to get a good night's rest a little of the time
3. Bladder symptoms interfered with your ability to get a good night's rest some of the time
4. Bladder symptoms interfered with your ability to get a good night's rest a good bit or most of the time
5. Bladder symptoms interfered with your ability to get a good night's rest all of the time

**Coping**
1. Bladder symptoms caused you to plan 'escape routes' to restrooms in public places none of the time
2. Bladder symptoms caused you to plan 'escape routes' to restrooms in public places a little of the time
3. Bladder symptoms caused you to plan 'escape routes' to restrooms in public places some of the time
4. Bladder symptoms caused you to plan 'escape routes' to restrooms in public places a good bit or most of the time
5. Bladder symptoms interfered with your ability to get a good night's rest all of the time

**Concern**
1. Bladder symptoms caused you embarrassment none of the time
2. Bladder symptoms caused you embarrassment a little of the time
3. Bladder symptoms caused you embarrassment some of the time
4. Bladder symptoms caused you embarrassment a good bit or most of the time
5. Bladder symptoms caused you embarrassment all of the time

**Table 3** Example question from the DCE surveys

| Health state A | Health state B |
|---|---|
| Bothered by an uncomfortable urge to urinate **a little bit or somewhat** | Bothered by an uncomfortable urge to urinate **a very great deal** |
| **Not at all** bothered by urine loss associated with a strong desire to urinate | Bothered by urine loss associated with a strong desire to urinate **a great deal** |
| Bladder symptoms interfered with your ability to get a good night's rest **none of the time** | Bladder symptoms interfered with your ability to get a good night's rest **some of the time** |
| Bladder symptoms caused you to plan 'escape routes' to restrooms in public places **none of the time** | Bladder symptoms caused you to plan 'escape routes' to restrooms in public places **some of the time** |
| Bladder symptoms caused you embarrassment **some of the time** | Bladder symptoms caused you embarrassment **a good bit or most of the time** |

Which health state do you think is better? (*please tick one box only*)

| A | B |
|---|---|
| | |

**Table 4** Characteristics of respondents in valuation surveys

| | *AQL-5D n (%)* | *AQL-5D postal survey N (%)* | *OAB-5D N (%)* | *OAB-5D postal survey* |
|---|---|---|---|---|
| Total | 307 | 263 | 311 | 402 |
| Age: | | | | |
| 18-25 | 34 (11.1%) | 9 (3.4%) | 37 (11.9%) | 14 (3.5%) |
| 26-35 | 57 (18.6%) | 35 (13.3%) | 57 (18.3%) | 47 (11.7%) |
| 36-45 | 61 (19.9%) | 45 (17.1%) | 61 (19.6%) | 71 (17.7%) |
| 46-55 | 50 (16.3%) | 56 (21.3%) | 51 (16.4%) | 81 (20.1%) |
| 56-65 | 45 (14.7%) | 64 (24.3%) | 45 (14.5) | 73 (18.2%) |
| >66 | 60 (19.5%) | 54 (20.5%) | 60 (19.3%) | 114 (28.4%) |
| Female | 168 (54.7%) | 148 (56.3%) | 160 (51.4%) | 236 (58.7%) |
| Married or living with partner | 214 (69.8%) | | 217 (69.8%) | |
| Experienced serious illness: | | | | |
| in family | 194 (63.4%) | | 176 (56.6%) | |
| in themselves | 94 (30.6%) | | 94 (30.2%) | |
| Degree or equivalent | 69 (22.5%) | | 85 (27.3%) | |
| Education after 17 | 140 (45.6%) | | 182 (58.5%) | |
| Renting property | 64 (20.8%) | | 63 (20.2) | |
| Found valuation tasks in interview difficult: | | | | |
| very difficult | 24 (7.9%) | | 13 (4.2%) | |
| quite difficult | 82 (26.7) | | 80 (25.9%) | |
| neither difficult nor easy | 52 (16.9) | | 70 (22.7%) | |
| Self-reported EQ-5D scores: | | | | |
| Male, female | 0.83, 0.84 | 0.81, 0.82 | 0.88, 0.88 | 0.87, 0.85 |

**Table 5** TTO and normalised rank and DCE model estimates[2] for AQL-5D

| *Dimension level* | *TTO* | *Rank*[2] | *Discrete choice experiment* | | | |
|---|---|---|---|---|---|---|
| | | | *Pooled data*[2] | *Warm data*[2] | *Cold data*[2] | *Pooled data normalised using TTO PITS* |
| concern2 | -0.028 | -0.018 | 0.012 | 0.021 | -0.006 | 0.008 |
| concern3 | -0.044* | -0.043* | -0.024 | -0.006 | -0.045 | -0.015 |
| concern4 | -0.054* | -0.092* | -0.099* | -0.101* | -0.103* | -0.058* |
| concern5 | -0.081* | -0.127* | -0.139* | -0.123* | -0.164* | -0.096* |
| breath2 | 0.000 | -0.038* | 0.025 | 0.044 | -0.010 | 0.025 |
| breath3 | -0.036* | -0.059* | -0.008 | 0.004 | -0.024 | -0.003 |
| breath4 | -0.101* | -0.068* | -0.116* | -0.092* | -0.153* | -0.057* |
| breath5 | -0.116* | -0.106* | -0.138* | -0.128* | -0.147* | -0.093* |
| pollution2 | -0.019 | -0.010 | 0.084* | 0.107* | 0.046 | 0.055* |
| pollution3 | -0.050* | -0.048* | -0.002 | 0.004 | -0.006 | 0.010 |
| pollution4 | -0.058* | -0.055* | -0.051* | -0.049 | -0.056 | -0.023 |
| pollution5 | -0.121* | -0.071* | -0.085* | -0.095* | -0.060 | -0.063* |
| sleep2 | 0.018 | -0.003 | -0.022 | -0.025 | -0.017 | -0.027 |
| sleep3 | 0.010 | -0.016 | -0.072* | -0.076* | -0.080 | -0.047* |
| sleep4 | -0.033* | -0.047* | -0.125* | -0.104* | -0.165* | -0.094* |
| sleep5 | -0.054* | -0.068* | -0.149* | -0.117* | -0.199* | -0.100* |
| activity2 | -0.039* | -0.064* | -0.056* | -0.064* | -0.051 | -0.032* |
| activity3 | -0.059* | -0.081* | -0.113* | -0.115* | -0.113* | -0.074* |
| activity4 | -0.175* | -0.163* | -0.247* | -0.262* | -0.232* | -0.158* |
| activity5 | -0.197* | -0.194* | -0.335* | -0.365* | -0.297* | -0.217* |
| Dead dummy | | -1.000* | -1.000* | -1.000* | -1.000* | |
| Number of observations | 2456 | 3041 | 2077 | 1336 | 741 | 1559 |
| Number of individuals | 307 | 306 | 263 | 168 | 95 | 263 |
| Inconsistencies[1] | 0 | 0 | 1 | 1 | 0 | 1 |
| No. predictions >0.05 from observed TTO | 19 | 24 | 34 | 33 | 39 | 24 |
| No. predictions >0.1 from observed TTO | 9 | 9 | 24 | 21 | 32 | 11 |
| MAD from TTO | 0.056 | 0.061 | 0.093 | 0.089 | 0.119 | 0.075 |
| RMSD from TTO | 0.070 | 0.079 | 0.118 | 0.111 | 0.149 | 0.093 |
| Mean Error | -0.025 | 0.001 | 0.059 | 0.036 | 0.102 | -0.060 |

*Notes:* *statistically significant at 5% level

[1] Relating to statistically significant dimensions only

[2] Adjusted Rank and DCE coefficients = estimated coefficient / dead dummy coefficient

**Table 6** TTO and normalised rank and DCE model estimates for OAB-5D

| *Dimension level* | *TTO* | *Rank*[2] | *Discrete choice experiment* | | | |
|---|---|---|---|---|---|---|
| | | | *Pooled data*[2] | *Warm data*[2] | *Cold data*[2] | *Pooled data normalised using TTO PITS* |
| urge2 | -0.033* | -0.065* | 0.048 | 0.072 | 0.034 | 0.024* |
| urge3 | -0.026* | -0.086* | 0.011 | 0.008 | 0.010 | 0.003 |
| urge4 | -0.065* | -0.119* | -0.109* | -0.117* | -0.106* | -0.035* |
| urge5 | -0.083* | -0.178* | -0.169* | -0.154* | -0.175* | -0.063* |
| urine2 | -0.018 | -0.028* | -0.023 | -0.056 | -0.012 | 0.002 |
| urine3 | -0.049* | -0.039* | -0.030 | 0.009 | -0.050 | -0.012 |
| urine4 | -0.030* | -0.060* | -0.134* | -0.061 | -0.171* | -0.043* |
| urine5 | -0.041* | -0.093* | -0.091* | -0.098* | -0.090* | -0.046* |
| sleep2 | -0.027* | -0.027* | 0.000 | -0.014 | 0.012 | -0.004 |
| sleep3 | -0.019 | -0.027* | 0.004 | -0.040 | 0.032 | -0.009 |
| sleep4 | -0.053* | -0.039* | -0.148* | -0.170* | -0.131* | -0.059* |
| sleep5 | -0.052* | -0.091* | -0.152* | -0.152* | -0.148* | -0.080* |
| coping2 | -0.004 | -0.011 | 0.087* | 0.117* | 0.074* | 0.002 |
| coping3 | -0.018 | -0.033* | -0.011 | 0.030 | -0.028 | -0.023* |
| coping4 | -0.021 | -0.040* | -0.009 | -0.008 | -0.011 | -0.028* |
| coping5 | -0.064* | -0.055* | -0.068* | -0.088* | -0.058 | -0.055* |
| concern2 | -0.031* | -0.036* | -0.028 | -0.029 | -0.027 | -0.018* |
| concern3 | -0.046* | -0.059* | -0.108* | -0.096* | -0.112* | -0.051* |
| concern4 | -0.085* | -0.095* | -0.235* | -0.244* | -0.231* | -0.095* |
| concern5 | -0.137* | -0.147* | -0.271* | -0.307* | -0.248* | -0.133* |
| Dead dummy | | -1.000* | -1.000* | -1.000* | -1.000* | |
| Number of observations | 2485 | 3040 | 3117 | 1050 | 2059 | 2347 |
| Number of individuals | 311 | 304 | 402 | 133 | 268 | 402 |
| Inconsistencies[1] | 3 | 0 | 2 | 2 | 2 | 1 |
| No. predictions >0.05 from observed TTO | 28 | 38 | 37 | 37 | 33 | 33 |
| No. predictions >0.1 from observed TTO | 5 | 18 | 29 | 29 | 31 | 14 |
| MAD from TTO | 0.061 | 0.068 | 0.112 | 0.120 | 0.112 | 0.086 |
| RMSD from TTO | 0.073 | 0.086 | 0.142 | 0.152 | 0.141 | 0.100 |
| Mean Error | -0.043 | 0.042 | 0.064 | 0.057 | 0.064 | -0.078 |

*Notes:* *statistically significant at 5% level

[1] Relating to statistically significant dimensions only

[2] Adjusted Rank and DCE coefficients = estimated coefficient / dead dummy coefficient

**Figure 1** Predictions of TTO, Rank and DCE models for AQL-5D in comparison to observed mean TTO

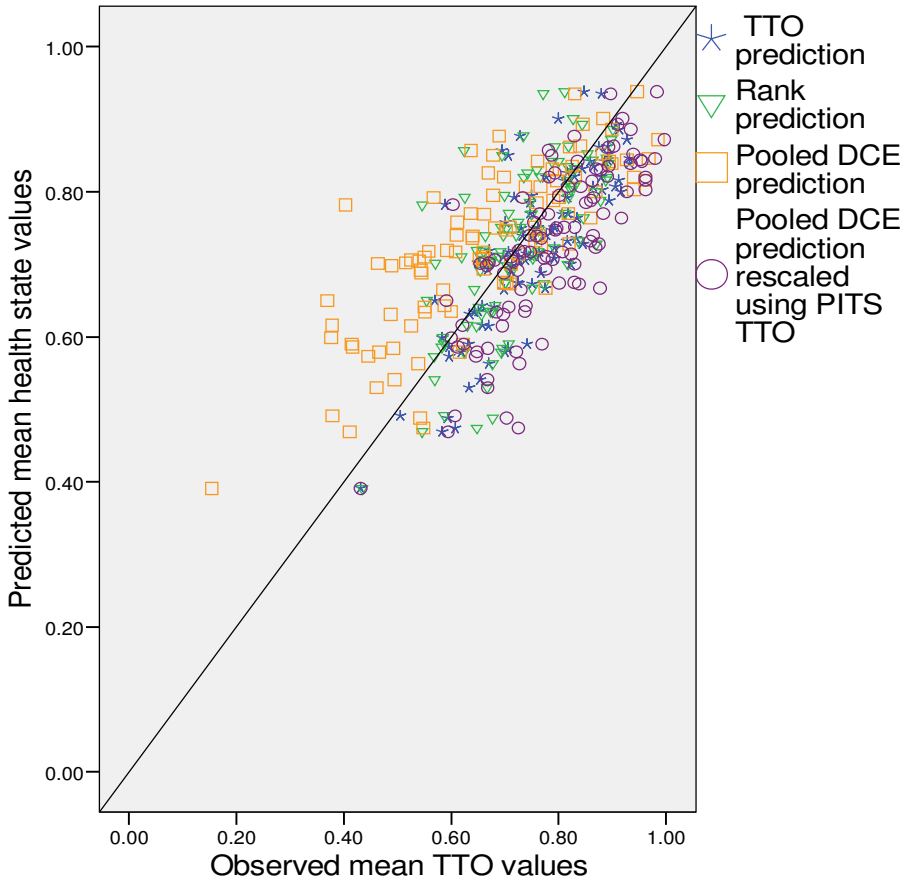

**Figure 2** Predictions of TTO, Rank and DCE models for OAB-5D in comparison to observed mean TTO